Overview of CLEF 2024 SimpleText Task 1

Retrieve Passages to Include in a Simplified Summary

Éric SanJuan (LIA, Avignon Université)
Stéphane Huet (LIA, Avignon Université)
Jaap Kamps (University of Amsterdam)
Liana Ermakova (HCTI, UBO)

CLEF 2024, Grenoble, France, September 9, 2024

Shared Tasks 2024

CLEF 2024 GRENOBLE

- *Task 1: Content Selection: retrieving passages to include in a simplified summary*
  - topical relevance
  - + text complexity scores (e.g., readability)
  - + authoritativeness scores (e.g., bibliometrics and altmetrics)
- *Task 2: Complexity Spotting*: identifying and explaining difficult concepts
  - difficult term detection *and* explanation
  - + usefulness of the provided explanation with regard to a query
  - + difficulty of the provided explanation
- *Task 3: Text Simplification*: simplify scientific text
  - expand the training and automatic evaluation data
  - + manual evaluation of information distortion & text complexity
  - + both sentence and passage level simplification
- *Task 4: SOTA?*: tracking the state-of-the-art in scholarly publications

Task 1: Content Selection

- *Task 1: Retrieving Passages to Include in a Simplified Summary*
- Given a popular science article targeted to a general audience, this task aims to retrieve passages that can help understand this article from a large corpus of academic abstracts and bibliographic metadata
- Citation Network Dataset: DBLP+Citation, ACM Citation network
  - 4,232,520 abstracts in English
- Topics based on 40 press articles + 114 manually extracted queries
  - 20 articles from *The Guardian*
  - 20 articles from *Tech Xplore*

Task 1: Examples

- Text of news articles as context (the **topic**)
  1. Patient data from GP surgeries sold to US companies
  2. Baffled by digital marketing? Find your way out of the maze
- Input: **query** based on these articles
  1 `patient data`
  2 `digital marketing`
  2 `advertising`
- Output:
  - Given the corpus of 4M articles (metadata+abstracts)
  - rank a list of abstracts relevant to the topic/query
  - in JSON format ($\sim$ trec_eval + passage)

The Guardian vs *Tech Xplore*

- G* queries : more ambiguous, social issues relating to IT
  - `Digital assistant`
  - `Biases`
  - `Drug discovery`
  - `Financial markets`
- T* queries: more technical, associated with a published scientific article
  - `RISC-V`
  - `OFDMA`
  - `photo transistor`
- Number of relevant documents
  - Some queries (e.g. `RNN`, `algorithm`, `system-on-chip`) are common keywords in DBLP (but retrieved documents have still to be associated with the specific topic)
  - Others are more original (e.g. `Crispr`, `nematode`), but have still relevant documents

Task 1: New queries

- 62 additional queries generated with OpenAI GPT 4 and post-edited
  - Only for *The Guardian* articles
  - Prompt: Find at least three topics in computer science in this paper
  - Query example: "how AI systems, especially virtual assistants, can perpetuate gender stereotypes?"
- Complete Open AI ChatGPT 4 run
  - Limitations: cannot access DBLP data, most of the provided references are out of the computer science field
  - Query delays too long to allow efficient interactive search.
- Towards Retrieval Augmented Generation combining DBLP search with Arxiv full-text content?

Task 1: Output format

CLEF 2024 GRENOBLE

- **run_id** Run ID starting with team ID, followed by task1 and run name
- **manual** Whether the run is manual {0,1}
- **topic_id** Topic ID
- **query_id** Query ID used to retrieve the document (if one of the queries provided for the topic was used; 0 otherwise)
- **doc_id** ID of the retrieved document (to be extracted from the JSON output)
- **rel_score** Relevance score (on the [0-1] scale)
- **comb_score** General score that may combine relevance and other aspects: readability, citation measures...
- **passage** Text of the selected passage

SimpleText'24 Submission Stats

Team	Task 1	Task 2			Task 3		Task 4		Total runs
		2.1	2.2	2.3	3.1	3.2	4.1	4.2	
AIIRLab	5	3	3		4	4			19
AMATU							3	9	12
Arampatzis	9	5	5	2	4	4			29
Elsevier	10				8	2			20
L3S							12	12	24
LIA	5								5
PiTheory					11	10			21
Sharigans	1	1	1		1	1			5
SINAI		3	3						6
SONAR					1				1
AB/DPV	1	1	1		1				4
Dajana/Katya		1			1				2
Frane/Andrea		1	1		1				3
Petra/Regina	1	1			1				3
Ruby	1	1			1	1			4
Tomislav/Rowan	2	2			1	1			6
UAmsterdam	6	1		2	4	6			19
UBO	1	1	1		2	2			7
UniPD		3	3						6
UZHPandas					11				11
Total runs	42	24	18	4	52	31	15	21	207

Participants approaches

- AB/DPV (1 run): ElasticSearch + FKGL
- Sharingans (1 run): ColBERT reranker + GPT3.5 to select passages
- Tomislav/Rowan (2 runs): ElasticSearch reranked using TF-IDF vectors + FKGL
- Petra/Regina (1 run) : ElasticSearch reranked using TF-IDF vectors + FKGL
- AIIRLab (5 runs): bi-encoder or a cross-encoder reranker, LLaMa3 as a pairwise reranker
- UBO (1 run): MonoT5 reranker
- UAmsterdam (6 runs): cross-encoder rerankings + filtering with FKGL
- Elsevier (10 runs): cross-encoder rerankers fine-tuned on a set of unlabeled scientific + generation new search queries with GPT-3.5
- LIA (5 runs): ElasticSearch + 4 extra baselines

Extra baselines

- Three bag-of-words models: ElasticSearch 7, MeiliSearch (bucket search) and boolean Search (PostgreSQL GIN text indexing) based on sparse vector document representation.
- Two MS MARCO Mini LM runs based on embedding vectors and dot product between the query and the abstract (vir_abstract) or the title (vir_title) using the pg_vector PostgreSQL extension and ivvflat dense vector index (k-means vector clustering with $\sqrt{|D|}$ centroids).
- API: https://guacamole.univ-avignon.fr/stvir_test?
  - **corpus**=[abstract|title]
  - **phrase**=varchar[300]
  - **length**=integer < 1000

Task 1: Evaluation

Qrels	Topics	#Queries	#Assessed abstracts			#Avg Ass.
			0	1	2	
2022 test	G1–G20, some T*	72	192	187	107	6.8
2023 train	G01–G15	29	728	338	237	44.9
2023 test	G16–G20, T01-T05	34	2260	357	1218	112.8
2024 train	G01–G20, T01-T05	64	3,675	768	1,655	95.5
2024 test	**G1.C1–G10.C1, T06–T11**	**30**	**2,775**	**1,500**	**579**	**128.5**
2024 test ext.	G1-G10, T01-T20	96	6,463	2,491	1,036	104.1

- Train data for system development:
  - 25 topics (mainly from *The Guardian*), with 64 specific queries.
- Test data:
  - Focus on queries out of the train data
  - Judgments on top 10 abstracts retrieved by all runs
- Evaluation measures:
  - Traditional IR metrics (relevance): NDCG, MAP...
  - Additional complexity/credibility aspect evaluation (automatic metrics)

Task 1: 2023 Results on Test Data

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
ElsevierSimpleText_run8	0.8082	0.5618	0.3515	0.5881	0.4422	0.2371	0.1633
ElsevierSimpleText_run7	0.7136	0.5618	0.4103	0.5704	0.4627	0.2626	0.1915
maine_CrossEncoder1^{rel}	0.8106	0.5382	0.4456	0.5675	0.4908	0.3317	0.2810
maine_CrossEncoderFinetuned1^{rel}	0.7691	0.5559	0.4441	0.5542	0.4840	0.3433	0.2572
maine_CrossEncoder1^{comb}	0.7309	0.5265	0.4500	0.5455	0.4841	0.3337	0.2754
ElsevierSimpleText_run5	0.6600	0.4765	0.3838	0.4826	0.4186	0.2542	0.1828
UAms_CE100^{rel}	0.7050	0.4912	0.4044	0.4782	0.4236	0.2616	0.2011
UAms_CE1k_Filter	0.6403	0.4765	0.3559	0.4533	0.3743	0.2727	0.1936
UAms_CE1k^{rel}	0.6329	0.4735	0.4044	0.4448	0.4049	0.2797	0.2051
Elastic baseline	0.6424	0.4059	0.3456	0.3910	0.3541	0.2501	0.1895
unimib_DoSSIER_2	0.5201	0.2853	0.2515	0.2980	0.2683	0.1898	0.1141
unimib_DoSSIER_4	0.5202	0.2853	0.2441	0.2972	0.2632	0.1873	0.1111
run-LIA.bm25	0.4536	0.1912	0.1338	0.2192	0.1700	0.1384	0.0515
run-LIA.all-MiniLM-L6-v2.query	0.3505	0.2000	0.1662	0.2019	0.1767	0.1956	0.0667
run-LIA.all-MiniLM-L6-v2.query-topic	0.3655	0.1765	0.1485	0.1912	0.1647	0.2043	0.0591

- Neural rankers outcompete lexical systems by a large margin
- In particular precision gains, some also recall
- Some submissions prioritized other aspects than relevance

2024 Results on Train Data (G01-G20 and T01-T05)

CLEF 2024 GRENOBLE

TURN IT SIMPLE

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder	0.7570	0.6467	0.4133	0.4955	0.4206	0.3463	0.2227
AIIRLab_Task1_LLaMAReranker2	0.7531	0.6200	0.4008	0.4708	0.4014	0.3364	0.2086
LIA_vir_title	0.6680	0.4433	0.2758	0.3405	0.2766	0.2742	0.1191
Arampatzis_1.GPT2_search_results	0.5732	0.3933	0.1967	0.2972	0.2184	0.0876	0.0676
UAms_Task1_Anserini_rm3	0.5613	0.3817	0.2833	0.2805	0.2541	0.2842	0.1408
Elsevier@SimpleText_task_1_run8	0.6173	0.3633	0.2458	0.2800	0.2406	0.1673	0.0993
LIA_vir_abstract	0.6015	0.3867	0.2633	0.2795	0.2405	0.2738	0.1168
LIA_bool	0.5646	0.3517	0.2400	0.2552	0.2238	0.2134	0.1037
Ruby_Task_1	0.5231	0.3050	0.2425	0.2387	0.2281	0.1696	0.1018
Elsevier@SimpleText_task_1_run10	0.5072	0.2983	0.2000	0.2335	0.1983	0.1356	0.0815
LIA_elastic	0.4540	0.2817	0.2067	0.2213	0.1977	0.2275	0.1103
AB/DPV_SimpleText_task1_FKGL	0.4538	0.2817	0.2067	0.2213	0.1977	0.1623	0.0948
Tomislav/Rowan_SimpleText_T1_1	0.5023	0.2683	0.1933	0.2108	0.1910	0.0972	0.0650
LIA_meili	0.4372	0.2883	0.1792	0.1833	0.1570	0.2024	0.0691
UBO_Task1_TFIDFT5	0.4134	0.1933	0.1775	0.1621	0.1625	0.1647	0.0730
Sharingans_Task1_marco-GPT3	0.4167	0.0417	0.0208	0.0658	0.0466	0.0085	0.0085
Tomislav/Rowan_SimpleText_T1_2	0.0108	0.0100	0.0067	0.0057	0.0051	0.0030	0.0011
Petra/Regina_results_simpleText_task_1	0.0013	0.0000	0.0025	0.0000	0.0018	0.0016	0.0004

2024 Results on Test Data (G01.C1-G10.C1 and T06-T11)

CLEF 2024 GRENOBLE
TURN IT SIMPLE

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9444	0.8167	0.5517	0.6311	0.5240	0.3559	0.2304
LIA_vir_title	0.8454	0.6933	0.4383	0.5090	0.4010	0.3594	0.1534
LIA_vir_abstract	0.7683	0.6000	0.4067	0.4269	0.3539	0.3857	0.1603
UAms_Task1_Anserini_rm3	0.7878	0.5700	0.4350	0.3945	0.3506	0.4010	0.1824
Arampatzis_1.GPT2_search^{rel}	0.6986	0.5100	0.2550	0.3522	0.2465	0.0742	0.0577
UBO_Task1_TFIDFT5	0.7132	0.4833	0.3817	0.3506	0.3215	0.2354	0.1274
LIA_bool*	0.7242	0.5233	0.3633	0.3409	0.2906	0.2661	0.1199
Elsevier@SimpleText_task_1_run8	0.7123	0.4533	0.3367	0.3152	0.2755	0.1582	0.0906
LIA_elastic	0.6173	0.3733	0.2900	0.2818	0.2442	0.3016	0.1325
AB&DPV_SimpleText_task1_FKGL^{rel}	0.6173	0.3733	0.2900	0.2818	0.2442	0.1966	0.1078
Ruby_Task_1^{rel}	0.5470	0.4233	0.3533	0.2790	0.2688	0.1980	0.1110
LIA_meili	0.6386	0.4700	0.2867	0.2736	0.2242	0.2377	0.0833
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5444	0.3733	0.2750	0.2477	0.2201	0.0963	0.0601
Sharingans_Task1_marco-GPT3	0.6667	0.0667	0.0333	0.1167	0.0807	0.0107	0.0107
Petra&Regina_simpleText_task_1	0.0026	0.0000	0.0050	0.0000	0.0035	0.0031	0.0007

2024 Results on Test Data (G01.C1-G10.C1)

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9500	0.7600	0.5125	0.5546	0.4777	0.3150	0.1919
LIA_vir_title	0.8014	0.6100	0.3750	0.4043	0.3307	0.2793	0.0985
LIA_bool*	0.7613	0.5800	0.4175	0.3531	0.3194	0.3384	0.1452
LIA_meili	0.7017	0.6100	0.3800	0.3477	0.2929	0.3175	0.1145
UAms_Task1_Anserini_rm3	0.7150	0.5250	0.4075	0.3248	0.3078	0.3486	0.1463
LIA_vir_abstract	0.6774	0.4900	0.3025	0.3053	0.2537	0.3020	0.0906
Arampatzis_1.GPT2_search^{comb}	0.6588	0.4900	0.2450	0.3050	0.2237	0.0651	0.0476
Elsevier@SimpleText_task_1_run8	0.6780	0.4400	0.2950	0.2847	0.2424	0.1131	0.0614
UBO_Task1_TFIDFT5	0.6198	0.4500	0.3425	0.2774	0.2610	0.1911	0.0903
Ruby_Task_1^{rel}	0.5550	0.4100	0.3600	0.2546	0.2587	0.1677	0.0966
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5550	0.4000	0.3200	0.2467	0.2380	0.1125	0.0675
LIA_elastic	0.5163	0.3000	0.2325	0.2010	0.1851	0.2540	0.0988
AB&DPV_SimpleText_task1_FKGL^{rel}	0.5163	0.3000	0.2325	0.2010	0.1851	0.1589	0.0762
Sharingans_Task1_marco-GPT3	0.5000	0.0500	0.0250	0.0816	0.0589	0.0070	0.0070

2024 Results on Test Data (T06-T11)

Run	MRR	Precision		NDCG		Bpref	MAP
		10	20	10	20		
AIIRLab_Task1_LLaMABiEncoder^{rel}	0.9500	0.7600	0.5125	0.5546	0.4777	0.3150	0.1919
LIA_vir_title	0.8014	0.6100	0.3750	0.4043	0.3307	0.2793	0.0985
LIA_bool	0.7613	0.5800	0.4175	0.3531	0.3194	0.3384	0.1452
LIA_meili	0.7017	0.6100	0.3800	0.3477	0.2929	0.3175	0.1145
UAms_Task1_Anserini_rm3	0.7150	0.5250	0.4075	0.3248	0.3078	0.3486	0.1463
LIA_vir_abstract	0.6774	0.4900	0.3025	0.3053	0.2537	0.3020	0.0906
Arampatzis_1.GPT2_search^{comb}	0.6588	0.4900	0.2450	0.3050	0.2237	0.0651	0.0476
Elsevier@SimpleText_task_1_run8	0.6780	0.4400	0.2950	0.2847	0.2424	0.1131	0.0614
UBO_Task1_TFIDFT5	0.6198	0.4500	0.3425	0.2774	0.2610	0.1911	0.0903
Ruby_Task_1^{rel}	0.5550	0.4100	0.3600	0.2546	0.2587	0.1677	0.0966
Tomislav/Rowan&Rowan_SimpleText_T1_1^{rel}	0.5550	0.4000	0.3200	0.2467	0.2380	0.1125	0.0675
LIA_elastic	0.5163	0.3000	0.2325	0.2010	0.1851	0.2540	0.0988
AB&DPV_SimpleText_task1_FKGL^{rel}	0.5163	0.3000	0.2325	0.2010	0.1851	0.1589	0.0762
Sharingans_Task1_marco-GPT3	0.5000	0.0500	0.0250	0.0816	0.0589	0.0070	0.0070

Evaluation of complexity and credibility (all 176 queries)

Run	Avg #Refs	Avg size of vocabulary	Ratio of long words	Ratio of complex words	FKGL	
					avg	median
AIIRLab_Task1_LLaMABiEncoder^{rel}	8.7	95.8	0.375	0.485	15.3	15.1
AIIRLab_Task1_LLaMAReranker2^{comb}	8.6	93.9	0.378	0.489	15.5	15.3
AIIRLab_Task1_LLaMAReranker2^{rel}	8.6	94	0.376	0.487	15.3	15.1
Arampatzis_1.GPT2_searchs	10.5	91.9	0.392	0.511	15.7	15.1
Elsevier@SimpleText_task_1_run4	10.7	99.1	0.375	0.495	15.1	14.9
Elsevier@SimpleText_task_1_run8	10.3	94.4	0.387	0.504	15.5	15.3
LIA_elastic	9.2	92.9	0.384	0.505	15.3	15.1
LIA_vir_abstract	7.2	69.8	0.378	0.484	14.6	14.3
LIA_vir_title	9.8	90.4	0.372	0.483	15	14.7
Sharingans_Task1_marco-GPT3	9.8	59.8	0.373	0.436	15.5	15.5
Uams_Task1_Anserini_rm3	11.9	112.9	0.387	0.508	16.8	16
UAms_Task1_CE1K_CAR^{comb}	10.2	98.5	0.363	0.483	13.8	13.5
Uams_Task1_CE1K	10.8	101.4	0.387	0.499	15.9	15.4
UBO_Task1_TFIDFT5	10.3	99.2	0.386	0.498	15.4	15.2

- Runs targeting relevant and more accessible abstracts
  - Performing competitive on retrieval effectiveness
  - readability level from "university" to "high school"
  - → avoiding very complex (but relevant) abstracts

Task 1: Findings

- Scientific passage retrieval test collection constructed in 2022-2024
  - High pooling diversity
  - Reusable with limited pooling bias
- Almost all submissions based on neural rankers
  - Crossencoders and biencoders popular and very effective
  - Training on scientific text helps
  - Small set of labeled train data can lead to overfitting (use with caution)
  - Queries alone are too ambiguous (topics or original articles have to be taken into consideration)
- Promising results for runs prioritizing credibility/complexity
  - Possible to factor the text complexity into the ranking
  - Guide users to accessible content first, and more complex text later

Task 1: To do list

- the vir_baseline shows that there are relevant documents which do not have a complete abstract
  => shall we enrich the corpus?
- consider the citation graphs to improve document retrieval
  => SOTA task ?
- back to effective passage retrieval ?
- still enriching the q-rels as training corpus ...
- PostGreSQL container with all data and baselines (32 Go) :
  https://guacamole.univ-avignon.fr/pubiutdev/simpletext/clef_st1.tar.gz

Thanks!

Website : https://simpletext-project.com[1]

E-mail : contact@simpletext-project.com

Twitter : https://twitter.com/SimpletextW

Google group : https://groups.google.com/g/simpletext

[1] This research was funded, in whole or in part, by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01.

References

- Liana Ermakova, Eric SanJuan, Stéphane Huet, Hosein Azarbonyad, Giorgio Maria Di Nunzio, Federica Vezzani, Jennifer D'Souza, and Jaap Kamps. "Overview of the CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone". In: *CLEF'24: Proceedings of the 15th International Conference of the CLEF Association*. Lecture Notes in Computer Science. Springer, 2024.
- Eric SanJuan, Stéphane Huet, Jaap Kamps, and Liana Ermakova. "Overview of the CLEF 2024 SimpleText Task 1: Retrieve Passages to Include in a Simplified Summary". In: *Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum*. CEUR Workshop Proceedings. CEUR-WS.org, 2024.